S9.6

# A NEW MIXED EXCITATION LPC VOCODER

Alan V. McCree and Thomas P. Barnwell III
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332, U.S.A.

ABSTRACT

The speech output of traditional pitch-excited LPC vocoders is of limited quality due to failures of the binary voicing decision, resulting in annoying buzzes, thumps, and tonal noises. This work introduces a new synthesizer structure for an LPC vocoder which increases the clarity and naturalness of the output speech. This synthesizer enhances the usual excitations of either periodic pulses or white noise by allowing pulse/noise mixtures and aperiodic pulses, and thus can generate a wider range of possible speech signals. The control algorithms for this new model replaces the traditional binary voicing decision with more robust periodicity, peakiness, and power level detectors, without significant increase in bit rate. As a result, the vocoder produces synthetic speech which is free of the usual LPC synthesis artifacts, even at bit rates below 2400 bps.

## 1 INTRODUCTION

The logical way to extend the successful Code Excited Linear Predictive (CELP) speech coder down to low bit rates (less than 3000 bits per second) is by drastically reducing the bit rate of the excitation signal. In the limiting case of a pitch-excited LPC vocoder, the excitation is parameterized as either a periodic impulse train or white noise. Unfortunately, these vocoders often produce speech which suffers from a strong synthetic quality, making them unacceptable for many applications. Male speakers often sound buzzy, while females sound mechanical and tense. In addition, occasional voicing decision errors result in annoying thumps and tonal noises. These problems stem from the inability of a simple pulse train to reproduce all kinds of voiced excitation, and from the inherent difficulty of correctly classifying natural human speech as either fully voiced or unvoiced.

This paper describes a new LPC vocoder synthesizer model. This new model has more freedom in the way it can parameterize the excitation signal and it is able to produce more realistic variations in voiced speech output. The new synthesizer allows a mixture of pulse and noise excitation, as well as aperiodic pitch pulses. Also, the algorithms to control this model are more reliable than the traditional binary voicing decision. Sections 2 and 3 of this paper describe the new synthesis structure and control, Section 4 discusses some related experiments on human speech perception, and Section 5 discusses the subjective evaluation of a prototype implementation of the new vocoder.

## 2 MIXED EXCITATION

### 2.1 Spectrally Flat Mixture Generation

The first feature of our new synthesizer model is the use of simultaneous pulse and noise excitation. The structure chosen in this work is motivated by a simple model for mixed excitation in natural human speech. If weakly voiced sounds are partially whispered, then the vocal tract is excited by both glottal pulse and noise simultaneously. Therefore, the LPC synthesizer could be excited by a sum of synthetic glottal pulses and white noise. However, this implies a low pass excitation signal, and in an LPC vocoder we must use a spectrally flat excitation.

We desire a structure which combines a low pass filtered pulse train with high pass filtered noise to produce a spectrally flat output regardless of the mixture ratio. Fortunately, this can be done using two simple first order FIR filters in the structure shown in Figure 1. The Appendix contains a derivation of the appropriate coefficient values as a function of the relative noise gain in the mixture. This mixture of low pass filtered pulses and high pass filtered noise is similar to that used by previous workers [1] [2], but it provides a spectrally flat output with continuously variable cutoff frequency using relatively little computation. The more complex multiband structure of Griffin and Lim [3] allows more freedom in controlling the noise spectrum, but this should only be necessary in an acoustically noisy environment, where the noise does not have a high pass "whisper" spectrum.

### 2.2 Mixture Control Algorithm

It is tempting to use periodicity indicators such as autocorrelation strength or power spectrum harmonic separation



CH2977-7/91/0000-0593 $1.00 © 1991 IEEE

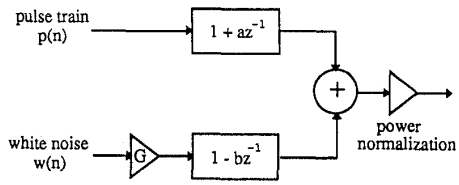


Figure 1. Mixed Excitation Structure

to control the excitation mixture, but other researchers [1] have reported limited success with this approach. During intervals of rising, falling, or jittery pitch or quick formant variations, the speech signal will appear to be less periodic, but it may still be strongly voiced. A control algorithm based on periodicity may put too much noise in the mixture in these regions, resulting in synthesized speech which sounds rough and hoarse.

We have found a simple but surprisingly effective alternative is to vary the pulse/whisper mixture with the relative power of the speech signal. In our current implementation, the synthesizer compares the interpolated power of each voiced pitch period to an adaptive estimate of fully voiced speech power, and adjusts the percentage of pulse and noise in the mixture according to this degree of voicing. If the current power is within 6 dB of fully voiced, the synthetic speech is strongly voiced (80% mixture). If the current power is more than 18 dB below fully voiced, the synthetic speech is weakly voiced (50% mixture). At intermediate power levels, the mixture percentage is linearly interpolated between these two extremes. For each voiced frame, the fully voiced power level is estimated as the maximum of the current voiced power and the previous fully voiced power decayed exponentially with a time constant of 1.3 seconds. Figure 2 shows the degree of voicing output from this algorithm for a typical sentence.

This algorithm ensures that emphasized vowels will be clear and noise-free, while feeble marginal voiced sounds are no longer artificially perfect. As a result, the buzzy quality is effectively eliminated without adding any hoarseness. Additional advantages of this approach are that it requires no increase in overall bit rate since the decision is made at the receiver, and it varies smoothly with time since the synthesizer can use the interpolated power level for each pitch period. However, the mixture control is used only for voiced sounds, and a binary voiced/unvoiced decision (voicing detector) is still required at the transmitter.

## 3 THE VOICING DECISION

In a traditional LPC vocoder, no algorithm has been found which yields a voicing decision which is always correct, so optimizing performance requires a tradeoff between the increased "buzz" caused by voicing too many frames and an increase in "thumps" due to erroneous noise bursts within voiced segments. With the mixed excitation synthesizer, the voicing detector can be made more sensitive without introducing buzz, but a new distortion soon becomes apparent. This is the presence of short isolated tones in the synthesized speech, especially for higher pitched female speakers. Since these tones cannot be eliminated by adding more high pass filtered noise to the mixture excitation, we presume that they result from undesirable periodicity in the lower harmonics of the excitation signal.

This periodicity can be removed by adding noise in the lower frequencies, but so much noise is required that the output speech sounds rough and occasional thumps are re-introduced. An alternative is to destroy the periodicity in the voiced excitation by varying each pitch period length with a pulse position jitter uniformly distributed up to ±25%. This allows the synthesizer to mimic the erratic glottal pulses which are often encountered in voicing transitions or in vocal fry [4]. This cannot be done for strongly voiced frames without introducing a hoarse quality, however, so a new control algorithm is needed to determine when the jitter should be added.

We have had good success by simply adding a third voicing state to the voicing decision which is made at the transmitter. The input speech is now classified as either voiced, jittery voiced, or unvoiced. In both voiced states, the synthesizer uses a mixed pulse/noise excitation, but in the jittery voiced state the synthesizer uses jittered, aperiodic pulses. This makes the problem of voicing detection easier, since strong voicing is defined by periodicity and is easily detected with any good periodicity detector. Jittery voicing corresponds to erratic glottal pulses, so it can be detected by either marginal periodicity or peakiness in the input speech. A peakiness detector will detect unvoiced plosives as well as jittery voicing, but this is not a problem since the use of randomly spaced pulses has previously been suggested to improve the synthesis for plosives [5]. In fact, misclassification of unvoiced frames as jittery voiced is almost always imperceivable, since the synthesizer uses aperiodic pulses with a noise mixture. Thus, the jittery voicing detector can be made very sensitive. Even the extreme case of synthesizing all unvoiced frames with the jittery voiced excitation produces good quality output speech, although sustained fricatives sound slightly unnatural.

This new LPC vocoder model avoids the traditional voicing decision tradeoff between buzz and thumps, and produces clean, natural sounding speech. These results also imply new information about human speech perception, which is discussed in the following section.

## 4 HUMAN EAR VOICING PERCEPTION

When the human ear decides that a synthesized speech segment is either buzzy or thumpy, it shows the ability to make a voicing decision based purely on the characteris-

tics of the excitation signal. The conventional LPC voiced excitation differs from the unvoiced in two ways: periodicity and pulsiness. We have conducted simple synthesis experiments which show that both of these properties are required in a fully voiced region. If either the periodicity is destroyed by introducing a uniformly distributed pulse position jitter up to ±25% or the pulsiness is destroyed by introducing large fixed phase differences between neighboring harmonics in a frequency domain synthesizer, the quality of voiced speech output is clearly degraded. Notice that we only require "bandpass pulsiness", since phase coherency is only needed within a fairly small band of frequencies, and broadband phase dispersion has little effect, presumably due to the bandpass filtering properties of the human ear.

We have also verified with informal experiments that neither of these properties is acceptable in fully unvoiced speech. If all unvoiced frames are synthesized with a periodic pulse train, a strong buzz is perceived. Interestingly, jittering the pulse positions does not diminish the buzz even though it destroys the periodicity. A mixed excitation of equal power low pass filtered impulse and high pass filtered noise eliminates the buzzy quality, however tones are now clearly evident due to pitch periodicity. If pulse positions are now jittered as above, the tonal quality also disappears and the synthesized speech is now of good quality. Similar results can be obtained from listening to the excitation signal directly, bypassing the LPC synthesis filtering.

From these experiments, we believe that the human ear is capable of separately detecting both periodicity and pulsiness. Both must be present for strongly voiced speech, but each results in distortion if present for all unvoiced speech. The buzz associated with LPC vocoders comes from excessive pulsiness in the higher frequencies, while excessive periodicity produces tonal noises.

## 5 IMPLEMENTATION AND EVALUATION

We have implemented an LPC vocoder with this new excitation model on a personal computer using the 'C' language with some inline assembler instructions. The system runs in real-time on a plug-in Banshee board from Atlanta Signal Processors, Inc. using a TMS320C30 microprocessor from Texas Instruments. Depending upon frame size and LPC order, the vocoder has a total bit rate of 2100 to 2800 bps. The LPC coefficients are determined by the autocorrelation technique over a Hamming window of length 25 msec, and quantized using line spectrum pair (LSP) differences [6]. The pitch is estimated from carefully normalized autocorrelations of the low pass filtered LPC residual signal. If the pitch correlation is strong, the frame is declared voiced. If the pitch correlation is of intermediate strength or if the fullband residual has a high peakiness value, the input is declared jittery voiced. Peakiness is defined by the ratio of the RMS power to the average value of the full-wave rectified residual [7]. An unvoiced frame is declared if none of these conditions are met. In addition, smoothing is applied to the pitch and voicing values using an algorithm which uses one past and one future frame. The synthesizer uses a uniform random number generator for unvoiced excitation, and impulses for voiced excitation. For jittery voiced frames, each pitch period length is multiplied by a random number uniformly distributed between 0.75 and 1.25. In both voiced states, the synthesizer uses a pulse/noise mixture with the relative noise gain determined by the transmitted speech power level.

In careful, but informal, listening tests, the new vocoder is clearly superior to a similar LPC vocoder with a binary voicing decision and no mixed excitation. In fact, the speech quality of this vocoder approaches that of the DoD standard 4800 bps CELP algorithm. A real-time demonstration will be available at the conference.

## 6 CONCLUSION

We have presented a new synthesizer structure for an LPC vocoder. By allowing separate control of periodicity and pulsiness in the excitation signal, this synthesizer can mimic a richer ensemble of possible speech characteristics. The control algorithms for this new model replace the traditional binary voicing decision with more reliable periodicity, peakiness, and power level detectors. As a result, the vocoder produces synthetic speech which is clean and natural, free of the usual LPC synthesis artifacts such as buzz, thumps, and tonal noises. A 2400 bps implementation of the new vocoder produces speech which is close in quality to that from the 4800 bps DoD CELP algorithm.

## REFERENCES


[1] J. Makhoul, R. Viswanathan, R. Schwartz, and A. W. F. Huggins, "A Mixed-Source Model for Speech Compression and Synthesis," *J. Acoust. Soc. Amer.*, vol. 64, pp. 1577–1581, Dec 1978.

[2] S. Y. Kwon and A. J. Goldberg, "An Enhanced LPC Vocoder with no Voiced/Unvoiced Switch," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, pp. 851–858, Aug 1984.

[3] D. W. Griffin and J. S. Lim, "Multiband Excitation Vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1223–1235, Aug 1988.

[4] W. Hess, *Pitch Determination of Speech Signals.* Springer, 1983.

[5] G. S. Kang and S. S. Everett, "Improvement of the Narrowband LPC Synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 1.7.1–1.7.4, 1984.

[6] J. Crosmer, *Very Low Bit Rate Speech Coding Using the Line Spectrum Pair Transformation of the LPC Coefficients.* PhD thesis, Georgia Institute of Technology, June 1985.

[7] D. L. Thomson and D. P. Prezas, "Selective Modeling of the LPC Residual During Unvoiced Frames: White Noise or Pulse Excitation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 3087–3090, 1986.

## APPENDIX: MIXTURE COEFFICIENT DERIVATION

This appendix contains a derivation of the low pass and high pass filter coefficient values needed to generate a spectrally flat mixture output as a function of the relative noise gain. The derivation is similar to that of Kwon and Goldberg [2], however this structure is simpler and results in a truly flat spectrum. We believe that their algorithm generates a low pass output spectrum.

In our structure, the pulse train $p(n)$ is low pass filtered by $H_p(z) = 1 + az^{-1}$, the noise $w(n)$ is high pass filtered by $H_w(z) = 1 - bz^{-1}$, $G$ is the noise gain, and the mixed excitation signal is given by the sum:

$$e(n) = G(h_w(n) * w(n)) + (h_p(n) * p(n)) \qquad (1)$$

Assuming the impulse train and the noise are white, independent, and of unity power spectrum, then the power spectrum of the mixed excitation is given by:

$$S_e(\omega) = G^2|H_w(\omega)|^2 + |H_p(\omega)|^2 \qquad (2)$$

$$S_e(\omega) = G^2(1 + b^2 - 2b\cos\omega) + 1 + a^2 + 2a\cos\omega \qquad (3)$$

Since we want a flat excitation spectrum, we set the frequency dependent terms to zero:

$$0 = G^2(-2b\cos\omega) + 2a\cos\omega \qquad (4)$$

$$a = bG^2 \qquad (5)$$

Generally, there is a range of possible values for the filter coefficients for any noise gain. We typically work with $G$ less than 1, $b$ equal to 1, and $a$ between 0 and 1. This always shapes the noise through a full differencing filter, thereby minimizing the low frequency components in the whisper noise.

Note that the spectrum of the pulse train is not really white, but consists only of harmonics of the pitch fundamental. This makes a rigorous definition of a "spectrally flat" mixture output difficult, but the concept is the same.

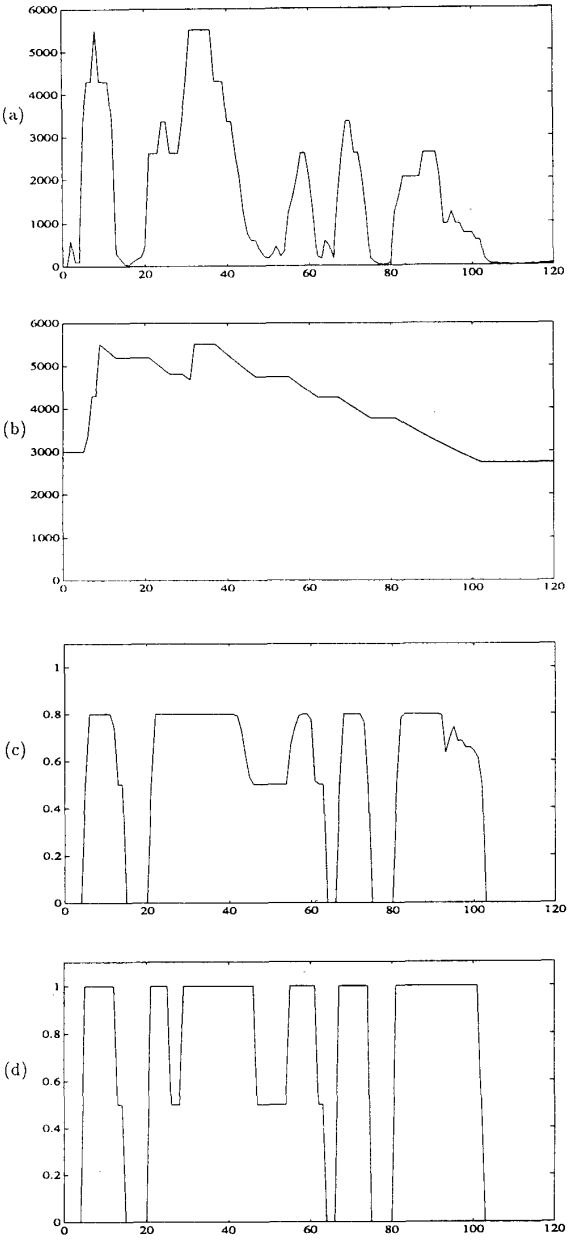


Figure 2. Voicing output for the sentence "Cats and dogs each hate the other" spoken by a male speaker: (a) quantized speech power, (b) fully voiced power estimate, (c) pulse mixture percentage, (d) voicing state (0 = unvoiced, 1 = voiced, 0.5 = jittery voiced).